

Standards Change Request

Use of Compression in PDS Archives

Elizabeth D. Rye

(modified)

SCR 3-1055

Sept. 13, 2005

Sept. 19, 2005

Problem:

As part of the effort to develop standards for the use of the JPEG 2000 compression format, it became immediately clear that with the exception of Zip Compression, there is no guidance whatsoever in the Standards Reference for those intending to compress archive data.

Proposed Solution:

This SCR attempts to establish a formal policy on the use of compression in PDS archives. The existing Zip Compression chapter of the Standards Reference will become one part of a new appendix on the specific rules for the implementation of individual compression formats. This SCR proposes to identify, but not populate, the other parts of the new appendix that will need to be populated in the future.

Requested Changes:

1. Approve the first attachment as a new PDS policy on data compression. (The place where this policy should be documented is TBD by the Management Council.)
2. Add a new appendix, "I", to the Standards Reference, entitled "Data Compression Formats". This appendix will contain one section for each compression format the PDS approves for the compression of archive data. Each section will contain a high level description of the compression format and a table of compression ratios for a common set of data files (to assist data providers in selecting the appropriate compression for their needs), any rules the PDS has established regarding the use of that particular format, a description of how a file utilizing that compression should be labeled, and a sample label. The sections for "CLEM-JPEG", "HUFFMAN FIRST DIFFERENCE", "PREVIOUS PIXEL", and "RUN LENGTH" will be populated, for the moment, with "TBDs". (These will need to be amended in the future through separate SCRs.)
3. Modify the current Zip Compression chapter to be the section of the above appendix that describes Zip compression, as shown in the second attachment.

Impact Assessment:

The impact of this SCR on PDS software and tools should be minimal, because it merely codifies existing practice that has never been explicitly stated. Of course, the Standards Reference will need to be updated as described above.

Additional Information:

None.

(This is a policy document, to be maintained either independently or as part of a yet to be written PDS policy document, depending on direction from the Management Council.)

Data Compression

In general, archiving data in a compressed format should be used sparingly. PDS recommends that data compression be used only in limited situations, such as:

1. To compress very large and infrequently used data, or *(exclusion of this statement is pending feedback from Simpson and Acton)*
2. To reduce the size of unmanageably large data files, or
3. To archive higher level data where the source (or lower level) product is readily available in a non-compressed PDS archive.

Note that compression of a particular data set must be approved by the relevant PDS discipline node.

PDS standards support two different approaches to data compression:

1. The first approach is to compress individual data objects using one of several supported methods (e.g., "Huffman first difference"). In this approach, the label describes the compressed file and the ENCODING_TYPE keyword indicates how the data object is to be decompressed by the user. The decompression software both decompresses the data object and creates a modified PDS label which describes the decompressed file.
2. In the second approach, an entire data file is compressed rather than a particular data object, using a commonly available utility (e.g., "Zip"). In this case, the PDS label describes both the compressed and decompressed files.

The individual compression schemes currently accepted by the PDS for compression of data are described in Appendix I of the PDS Standards Reference, along with the detailed labeling requirements for each.

The following rules apply to all data sets which include compressed data:

1. Compression formats must be approved for use by the PDS before they are included in an archive.
2. Only lossless compression may be used.
3. Product meta-data must include an indication of the version of the compression algorithm (or decompression software) used in compressing the product. (The `ENCODING_TYPE` and `SOFTWARE_VERSION_ID` keywords are two places this information may be stored.)
4. The PDS must have a copy of the specification or standard defining the compression algorithm used, at the version level that was used. If legally permitted, the documentation should be included in the archive.
5. The compressed products must be validated to comply with the specification or standard defining the compression algorithm used.
6. Decompression software must be capable of producing a correctly formatted and labeled decompressed PDS data file. Additional output formats are permitted. Source code and executables for decompression programs must be provided to the relevant PDS discipline node at the time an archive is delivered. It is recommended that these be included in the archive. Well documented decompression algorithms must be included in the archive.
7. The compression and decompression software must be validated on a number of test data files to verify that the input and output files are identical. Thereafter, a random sampling of data products in the archive should be decompressed as part of the validation process.
8. The compressed products, decompression algorithms, and decompression software must all be available for use by the PDS and its users on a royalty and license fee free basis.

Appendix I. Data Compression Formats

This appendix provides a brief description of each of the compression formats that has been approved by the PDS for archive data. For rules governing the use of data compression in PDS archives, please see the PDS policy on data compression.

Each section in this appendix includes a high level description of the compression format, a table showing compression ratios, PDS-specific implementation rules, and information about how to properly label files implementing the compression algorithm. Each section should also include a sample label.

The files to be used for the comparison of compression ratios are the following:

1. “Busy” Image – TBD
2. “Smooth” Image – TBD
3. ASCII Table – “sol06_eu.lbl” and “sol06_eu.tab” – This is a modified version of a Mars Pathfinder Rover product containing engineering data in an ASCII table. The table and label combined are roughly 10MB in size.

The above files are available on the PDS web site at the following URL:

<http://pds-engineering.jpl.nasa.gov/TBD>

Chapter Contents

Appendix I.	Data Compression Formats.....	I-1
I.1	CLEM-JPEG.....	I-3
I.2	HUFFMAN FIRST DIFFERENCE.....	I-4
I.3	PREVIOUS PIXEL	I-5
I.4	RUN LENGTH.....	I-6
I.5	ZIP.....	I-7

I.1 CLEM-JPEG

TBD

I.1.1 Table of Compression Ratios

“Busy” Image	“Smooth” Image	ASCII Table
TBD	TBD	N/A

I.1.2 PDS Implementation Rules

TBD

I.1.3 Labeling

TBD

I.1.4 Label Example

TBD

I.2 HUFFMAN FIRST DIFFERENCE

TBD

I.2.1 Table of Compression Ratios

“Busy” Image	“Smooth” Image	ASCII Table
TBD	TBD	N/A

I.2.2 PDS Implementation Rules

TBD

I.2.3 Labeling

TBD

I.2.4 Label Example

TBD

I.3 PREVIOUS PIXEL

TBD

I.3.1 Table of Compression Ratios

“Busy” Image	“Smooth” Image	ASCII Table
TBD	TBD	N/A

I.3.2 PDS Implementation Rules

TBD

I.3.3 Labeling

TBD

I.3.4 Label Example

TBD

I.4 RUN LENGTH

TBD

I.4.1 Table of Compression Ratios

“Busy” Image	“Smooth” Image	ASCII Table
TBD	TBD	N/A

I.4.2 PDS Implementation Rules

TBD

I.4.3 Labeling

TBD

I.4.4 Label Example

TBD

I.5 ZIP

The Zip method was chosen because the algorithm and supporting software for all major platforms are available without charge to the general user community. The *Info-Zip Consortium* and Info-Zip working group, for example, provide information and software at this URL:

<http://www.info-zip.org>

This same information is available on line from PDS at:

<http://pds.jpl.nasa.gov>

I.5.1 Table of Compression Ratios

“Busy” Image	“Smooth” Image	ASCII Table
TBD	TBD	27.7:1

I.5.2 PDS Implementation Rules

A volume containing zip files with combined-detached labels as presented below conforms to all established PDS standards *provided both the zip file and its constituent data files are archived*. The unique feature of a Zip-compressed PDS archive volume is that only the zip files appear; the UNCOMPRESSED_FILE objects described by the labels are not present on the volume, but can be obtained by unzipping the zip files provided.

In the interests of long-term archiving, a PDS archive zip file must include all the support files required to completely reconstitute the labeled data files. Specifically, the zipped archive must include not only the data files, but also the label file(s) for the uncompressed data. Ideally, any .FMT files referenced by ^STRUCTURE keywords in the labels should also be included in the zip file.

Note: These additional .LBL and .FMT files do not need to be described by UNCOMPRESSED_FILE objects in the label, because PDS label and format files never require labels. Furthermore, the sizes of these files do not need to be included in the value of the REQUIRED_STORAGE_BYTES keyword. However, the names of these files do need to be included in the list of UNCOMPRESSED_FILE_NAME values.

I.5.3 Labeling

When archiving data in Zip format, two files need to be considered: (1) the zip file itself, and (2) the data file produced by decompressing the zip file. PDS strongly recommends that these two files have the same name but different extensions: “.ZIP” for the zip file and a more descriptive extension (e.g., “.DAT” or “.IMG”) for the unzipped file. The “.ZIP” file extension is reserved exclusively for zip-compressed files within the PDS.

PDS does not recommend the practice of compressing multiple data files into a single zip file, unless those files reside in the same directory and have the same name, but different extensions. For example, if file “ABC.IMG” contains an image and file “ABC.TAB” contains a table of additional information relevant to that image, then both files can be archived in the file “ABC.ZIP”. This will minimize the potential confusion for a user who may not be able to locate a desired file because it is hidden inside a zip file with a different name.

Like all PDS data files, both the zipped and the unzipped data files require labels. Both files must be described by a single, detached PDS label file using the combined-detached label approach (see Section 5.2.2). Attached labels are not permitted for Zip-compressed data, because the user must be able to examine the label before deciding whether or not to decompress the file. In a combined-detached label, each individual file is described as a FILE object. Here is the general framework:

```

PDS_VERSION_ID      = PDS3
DATA_SET_ID         = ...
PRODUCT_ID          = ...
    (other parameters relevant to both Zipped and Unzipped files)

OBJECT              = COMPRESSED_FILE
    (parameters describing the compressed file)
END_OBJECT          = COMPRESSED_FILE

OBJECT              = UNCOMPRESSED_FILE
    (parameters describing the first uncompressed file)
END_OBJECT          = UNCOMPRESSED_FILE

OBJECT              = UNCOMPRESSED_FILE
    (parameters describing a second uncompressed file, if present)
END_OBJECT          = UNCOMPRESSED_FILE
END
```

The first FILE object, the COMPRESSED_FILE, refers to the zipped file; additional FILE objects, called UNCOMPRESSED_FILES, refer to the decompressed data file(s) that the user will obtain by unzipping the first.

The zip file is described via a “minimal label” (see Section 5.2.3). The following keywords are required:

FILE_NAME	= name of the zipfile
RECORD_TYPE	= UNDEFINED
ENCODING_TYPE	= ZIP
INTERCHANGE_FORMAT	= BINARY
UNCOMPRESSED_FILE_NAME	= a list of the names of all the files archived in the zipfile
REQUIRED_STORAGE_BYTES	= approximate total number of bytes in the data files
DESCRIPTION	= a brief description of the zipfile format

Typically, the DESCRIPTION is given as a pointer to a file called “ZIPINFO.TXT” found in the DOCUMENT directory on the same volume.

The subsequent UNCOMPRESSED_FILE object(s) contain complete descriptions of the data files obtained by unzipping the zip file.

I.5.4 Label Example

The following is an example of a PDS label for a Zip-compressed data file.

PDS_VERSION_ID	= PDS3
DATA_SET_ID	= "HST-S-WFPC2-4-RPX-V1.0"
SOURCE_FILE_NAME	= "U2ON0101T.SHF"
PRODUCT_TYPE	= OBSERVATION_HEADER
PRODUCT_CREATION_TIME	= 1998-01-31T12:00:00
OBJECT	= COMPRESSED_FILE
FILE_NAME	= "0101_SHF.ZIP"
RECORD_TYPE	= UNDEFINED
ENCODING_TYPE	= ZIP
INTERCHANGE_FORMAT	= BINARY
UNCOMPRESSED_FILE_NAME	= {"0101_SHF.DAT", "0101_SHF.LBL"}
REQUIRED_STORAGE_BYTES	= 34560
^DESCRIPTION	= "ZIPINFO.TXT"
END_OBJECT	= COMPRESSED_FILE
OBJECT	= UNCOMPRESSED_FILE
FILE_NAME	= "0101_SHF.DAT"
RECORD_TYPE	= FIXED_LENGTH
RECORD_BYTES	= 2880
FILE_RECORDS	= 12
^FITS_HEADER	= ("0101_SHF.DAT", 1 <BYTES>)
^HEADER_TABLE	= ("0101_SHF.DAT", 25921 <BYTES>)
OBJECT	= FITS_HEADER
HEADER_TYPE	= FITS

```

    INTERCHANGE_FORMAT      = ASCII
    RECORDS                  = 7
    BYTES                    = 20160
    ^DESCRIPTION             = "FITS.TXT"
    END_OBJECT               = FITS_HEADER

    OBJECT                   = HEADER_TABLE
    NAME                     = HEADER_PACKET
    INTERCHANGE_FORMAT      = BINARY
    ROWS                     = 965
    COLUMNS                 = 1

    ROW_BYTES                = 2
    DESCRIPTION              = "This is the HST standard header packet
                                containing observation parameters.  It is
                                stored as a sequence of 965 two-byte
                                integers.  For more detailed information,
                                contact Space Telescope Science Institute."

    OBJECT                   = COLUMN
    NAME                     = PACKET_VALUES
    DATA_TYPE               = MSB_INTEGER
    START_BYTE               = 1
    BYTES                    = 2
    END_OBJECT               = COLUMN
    END_OBJECT               = HEADER_TABLE

    END_OBJECT               = UNCOMPRESSED_FILE
    END

```

I.5.5 ZIPINFO.TXT Example

While the ZIPINFO.TXT file is not required, it is strongly recommended that this file be included as part of the process of documenting the contents of a zip file. The following is an example ZIPINFO.TXT file and the type of information that should be included in the ZIPINFO.TXT file:

```

    PDS_VERSION_ID          = PDS3
    RECORD_TYPE              = STREAM

    OBJECT                   = TEXT
    PUBLICATION_DATE         = 1999-07-26
    NOTE                     = "This file provides an overview of the ZIP
                                file format."

    END_OBJECT               = TEXT
    END

```

Many of the files in this data set are compressed using Zip format. They are all indicated by the extension ".ZIP". ZIP is a utility that

compresses files and also allows for multiple files to be stored in a single Zip archive. You will need the UNZIP utility to extract the files.

The SOFTWARE directory on this volume contains a complete description of the Zip file format and also the complete source code for the UNZIP utility. The file format and file decompression algorithms are described in the file SOFTWARE/APPNOTE.TXT.

It is far simpler to obtain a pre-built binary of the UNZIP application for your platform. Binaries for most platforms are available from the Info-ZIP web site, currently at this URL:

<http://www.info-zip.org/>

The same information can also be found at the PDS Engineering Node's web site, currently at:

<http://pds.jpl.nasa.gov/>